CS540 Machine learning
L9 Bayesian statistics

Last time

- Naïve Bayes
- Beta-Bernoulli

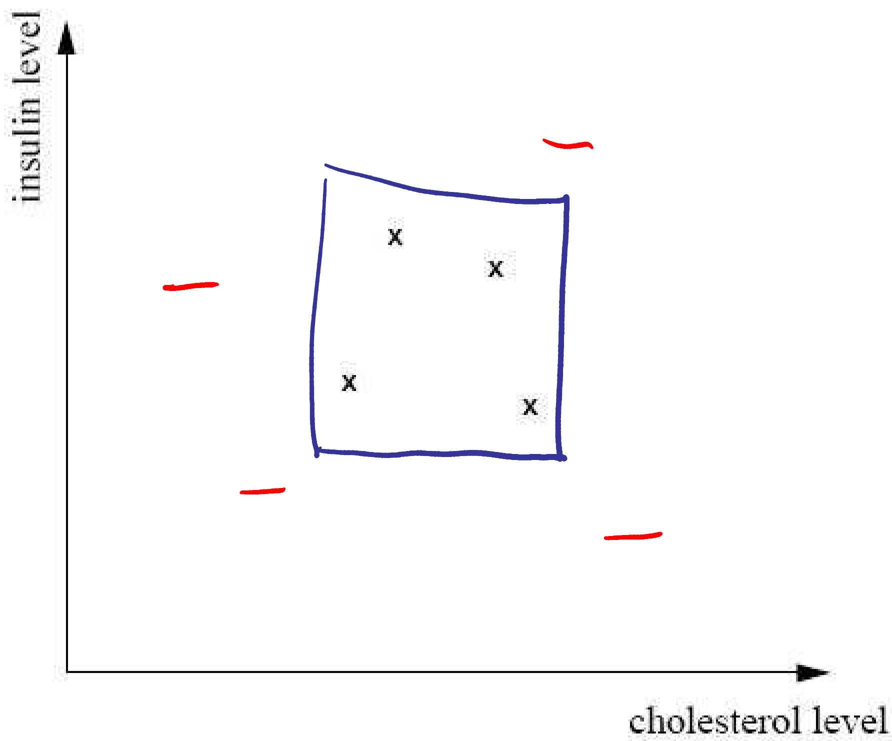
Outline

- Bayesian concept learning
- Beta-Bernoulli model (review)
- Dirichlet-multinomial model
- Credible intervals

Bayesian concept learning

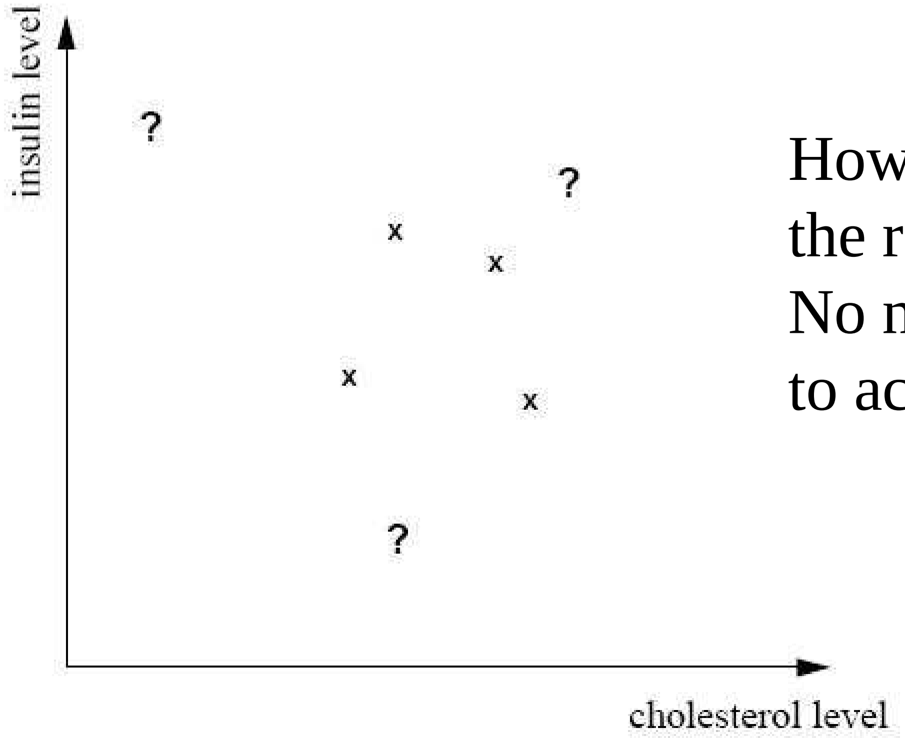
Based on Josh Tenenbaum's PhD thesis (MIT BCS 1999)

"Concept learning" (binary classification) from positive and negative examples



"healthy levels"

Concept learning from positive only examples

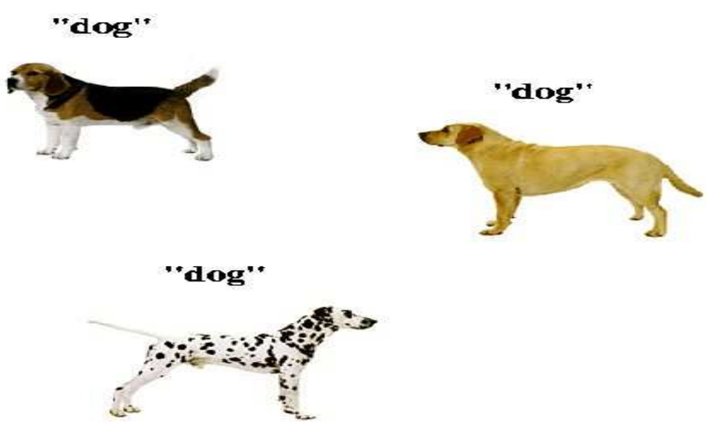


"healthy levels"

How far out should the rectangle go?
No negative examples to act as an upper bound.

Human learning vs machine learning/ statistics

- Most ML methods for learning "concepts" such as "dog" require a large number of positive and negative examples
- But people can learn from small numbers of positive only examples (look at the doggy!)
- This is called "one shot learning"



Everyday inductive leaps

How can we learn so much about . . .

- Meanings of words
- Properties of natural kinds
- Future outcomes of a dynamic process
- Hidden causal properties of an object
- Causes of a person's action (beliefs, goals)
- Causal laws governing a domain

. . . from such limited data?

The Challenge

- How do we generalize successfully from very limited data?
  - Just one or a few examples
  - Often only positive examples
- Philosophy:
  - Induction called a “problem”, a “riddle”, a “paradox”, a “scandal”, or a “myth”.
- Machine learning and statistics:
  - Focus on generalization from many examples, both positive and negative.

The solution: Bayesian inference

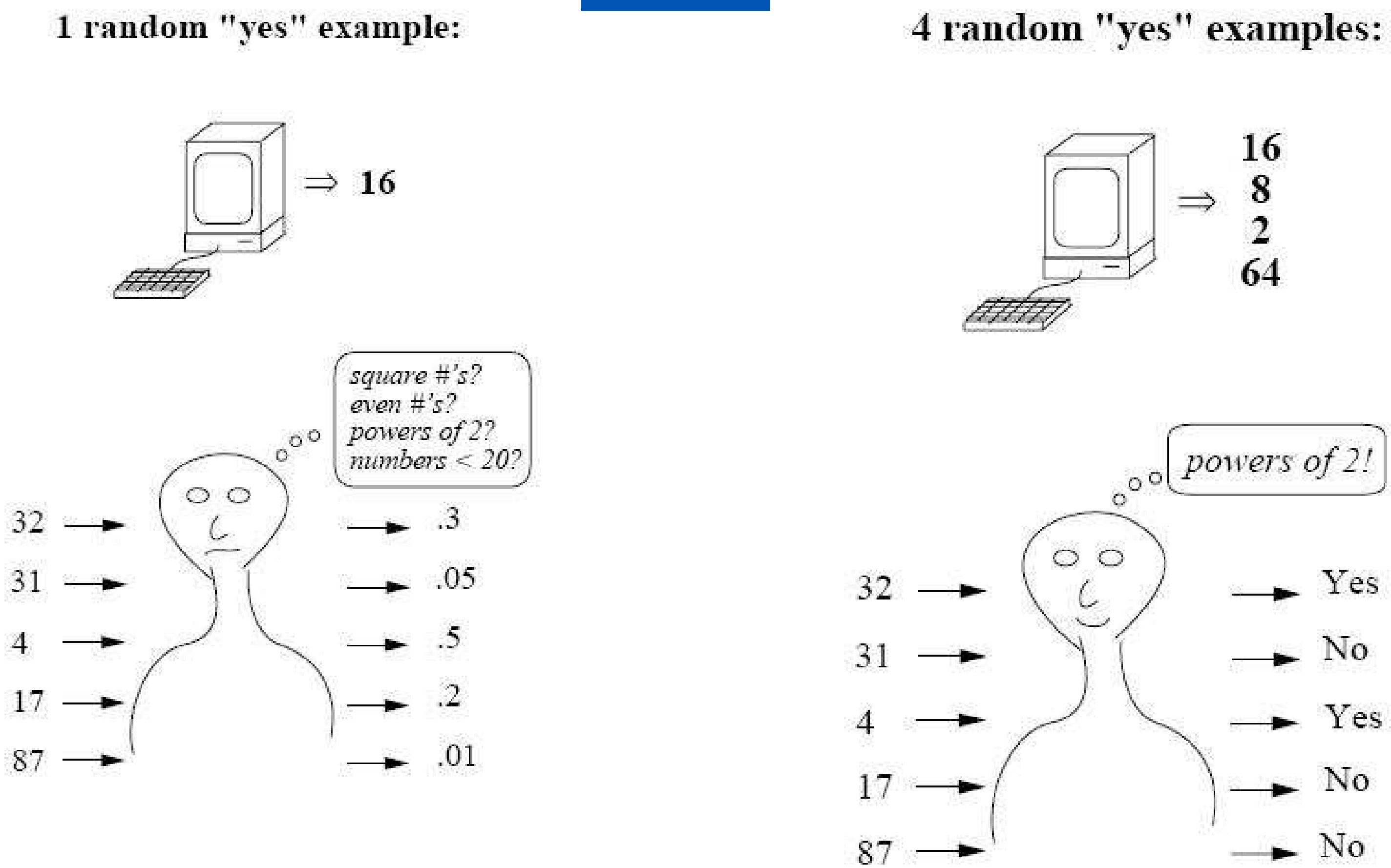
- Bayes' rule: $$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)}$$
- Various compelling (theoretical and experimental) arguments that one should represent one's beliefs using probability and update them using Bayes rule

Bayesian inference: key ingredients

- Hypothesis space H
- Prior p(h)
- Likelihood p(D|h)
- Algorithm for computing posterior p(h|D)

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

The number game



- Learning task:
  - Observe one or more examples (numbers)
  - Judge whether other numbers are "yes" or "no".

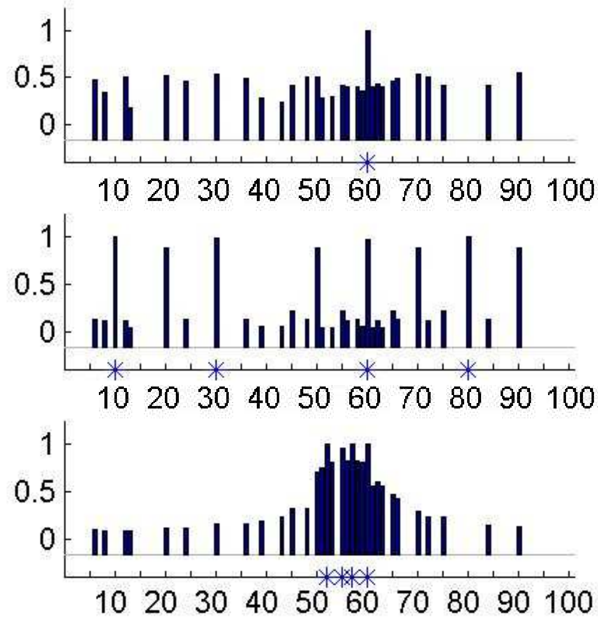
The number game

Examples of "yes" numbers	Hypotheses
60	multiples of 10 even numbers ? ? ?
60 80 10 30	multiples of 10 even numbers
60 63 56 59	numbers "near" 60

60

60 80 10 30

60 52 57 55



Diffuse similarity

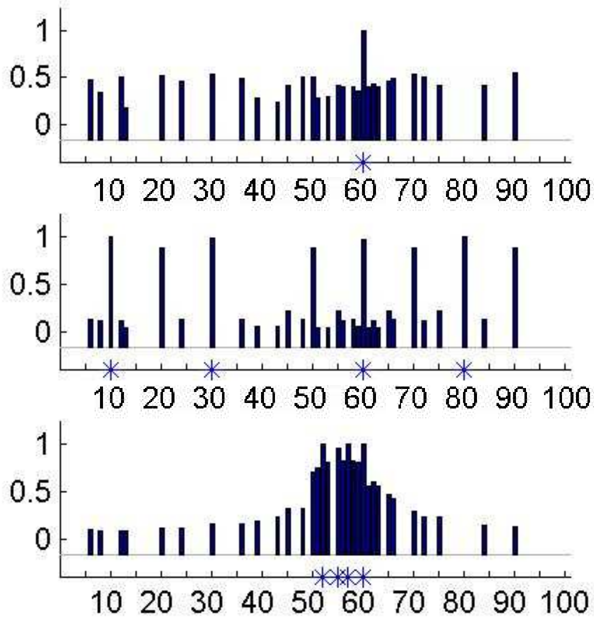
Rule:
"multiples of 10"

Focused similarity:
numbers near 50-60

60

60 80 10 30

60 52 57 55



Diffuse similarity

Rule:
"multiples of 10"

Focused similarity:
numbers near 50-60

Some phenomena to explain:

- People can generalize from just positive examples.
- Generalization can appear either graded (uncertain) or all-or-none (confident).

Bayesian model

- *H*: Hypothesis space of possible concepts:
- $X = \{x_1, \ldots, x_n\}$: *n* examples of a concept *C*.
- Evaluate hypotheses given data using Bayes' rule:

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$

  - $p(h)$ ["prior"]: domain knowledge, pre-existing biases
  - $p(X|h)$ ["likelihood"]: statistical information in examples.
  - $p(h|X)$ ["posterior"]: degree of belief that *h* is the true extension of *C*.

Hypothesis space

- Mathematical properties (~50):
  - odd, even, square, cube, prime, …
  - multiples of small integers
  - powers of small integers
  - same first (or last) digit
- Magnitude intervals (~5000):
  - all intervals of integers with endpoints between 1 and 100
- Hypothesis can be defined by its **extension**

$$h = \{x : h(x) = 1,\ x = 1, 2, \ldots, 100\}$$

Likelihood p(X|h)

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as *n* increases.

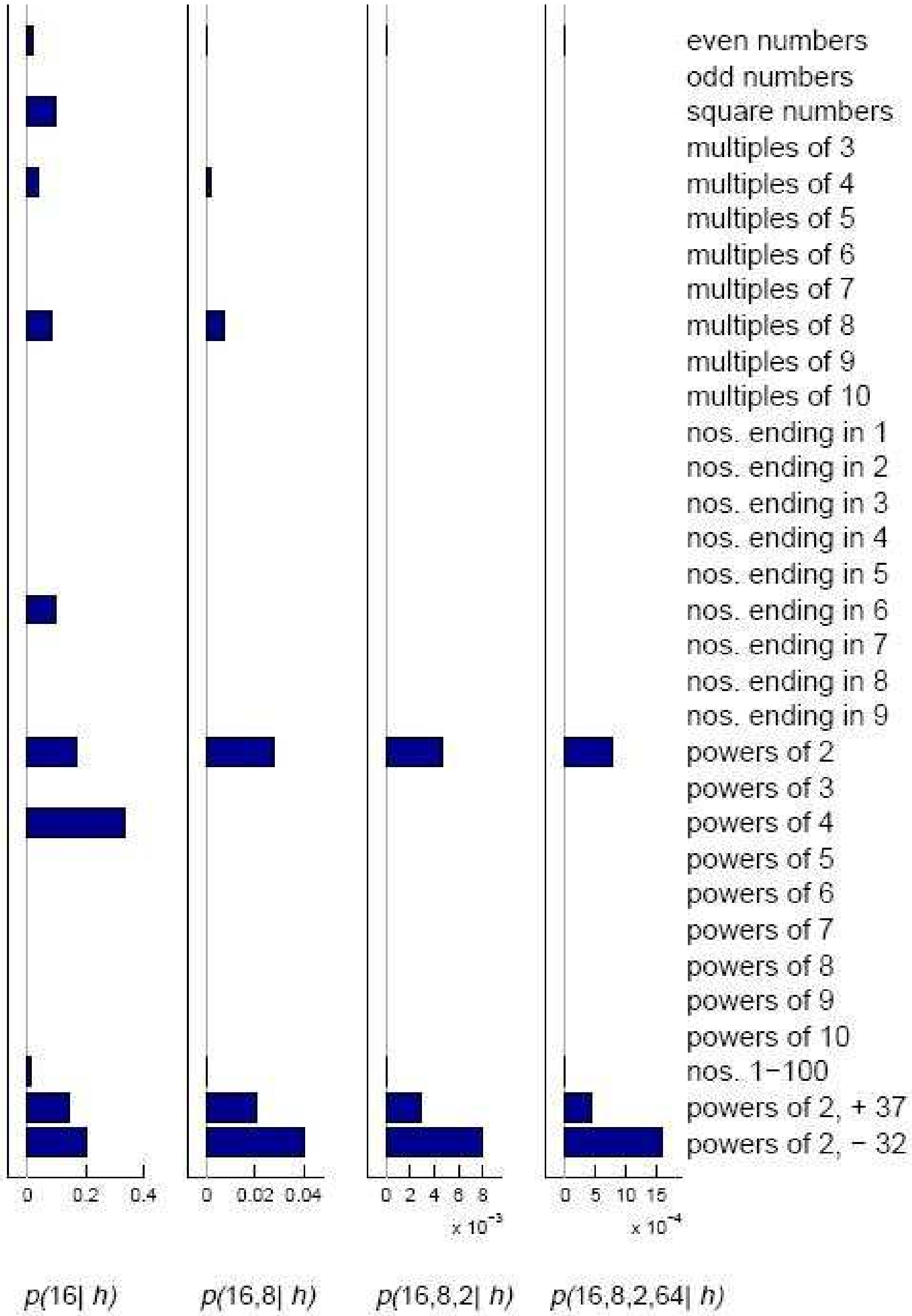
$$p(X \mid h) = \left[\frac{1}{\text{size}(h)} \right]^n \text{ if } x_1, \ldots, x_n \in h$$

$$= 0 \text{ if any } x_i \notin h$$

- Follows from assumption of randomly sampled examples (**strong sampling**).
- Captures the intuition of a representative sample.

Example of likelihood

- X={20,40,60}
- H1 = multiples of 10 = {10,20,…,100}
- H2 = even numbers = {2,4,…,100}
- H3 = odd numbers = {1,3,…,99}
- P(X|H1) = 1/10 * 1/10 * 1/10
- p(X|H2) = 1/50 * 1/50 * 1/50
- P(X|H3) = 0

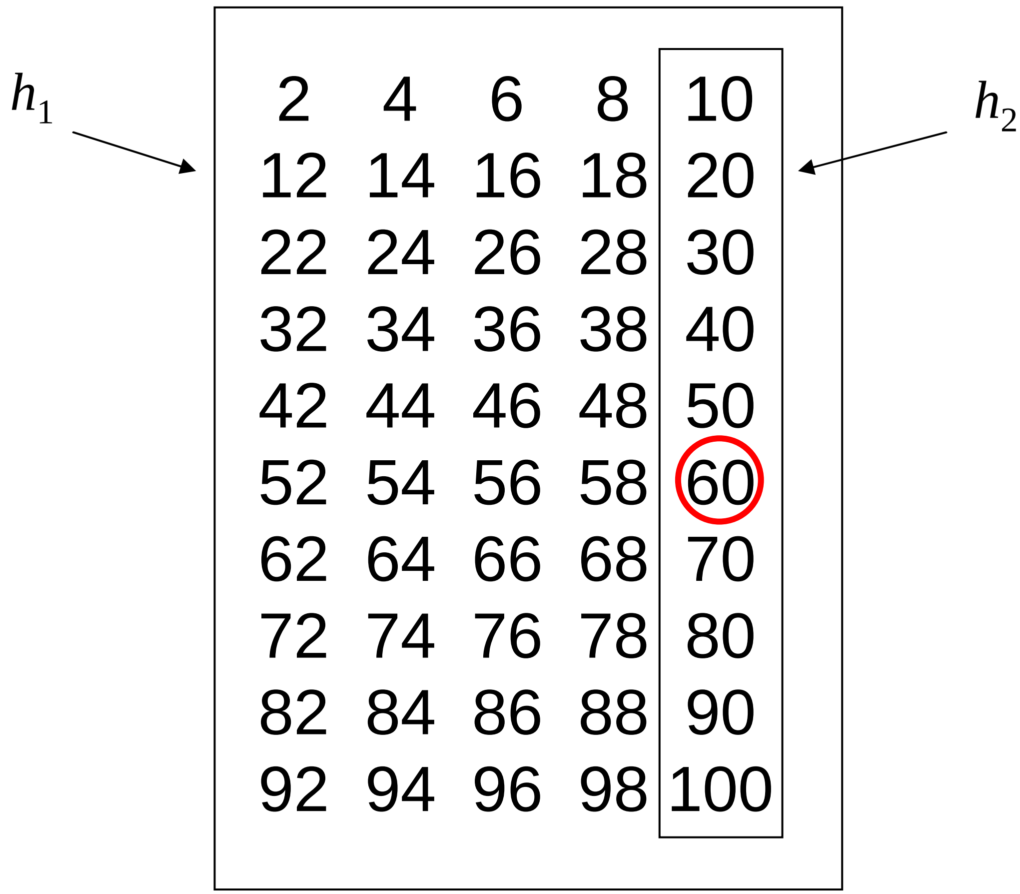

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0 0.2 0.4
0 0.02 0.04
0 2 4 6 8
x 10^-3
0 5 10 15
x 10^-4
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64| h)

Size principle

h_1 h_2

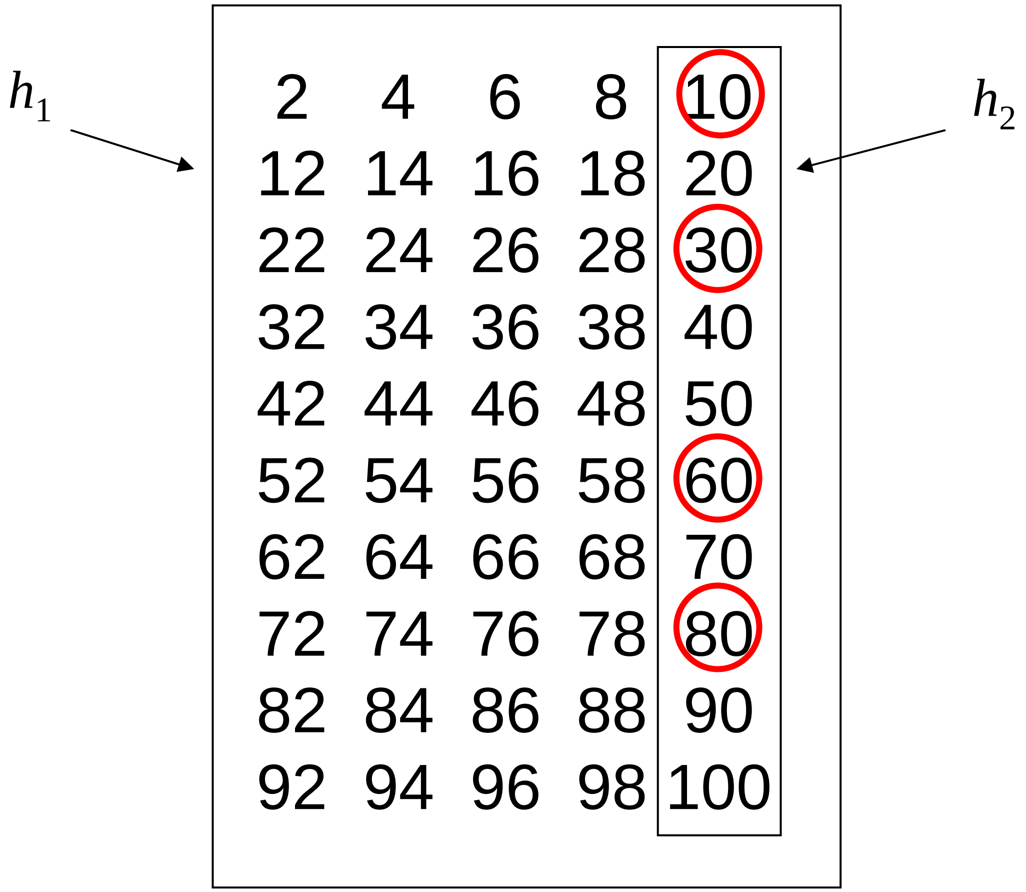
2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

Size principle



Data slightly more of a coincidence under h_1

Size principle



Data *much* more of a coincidence under h_1

Prior p(h)

- X={60,80,10,30}
- Why prefer "multiples of 10" over "even numbers"?
  - Size principle (likelihood)
- Why prefer "multiples of 10" over "multiples of 10 except 50 and 20"?
  - Prior
- Cannot learn efficiently if we have a uniform prior over all 2^{100} logically possible hypotheses

Need for prior (inductive bias)

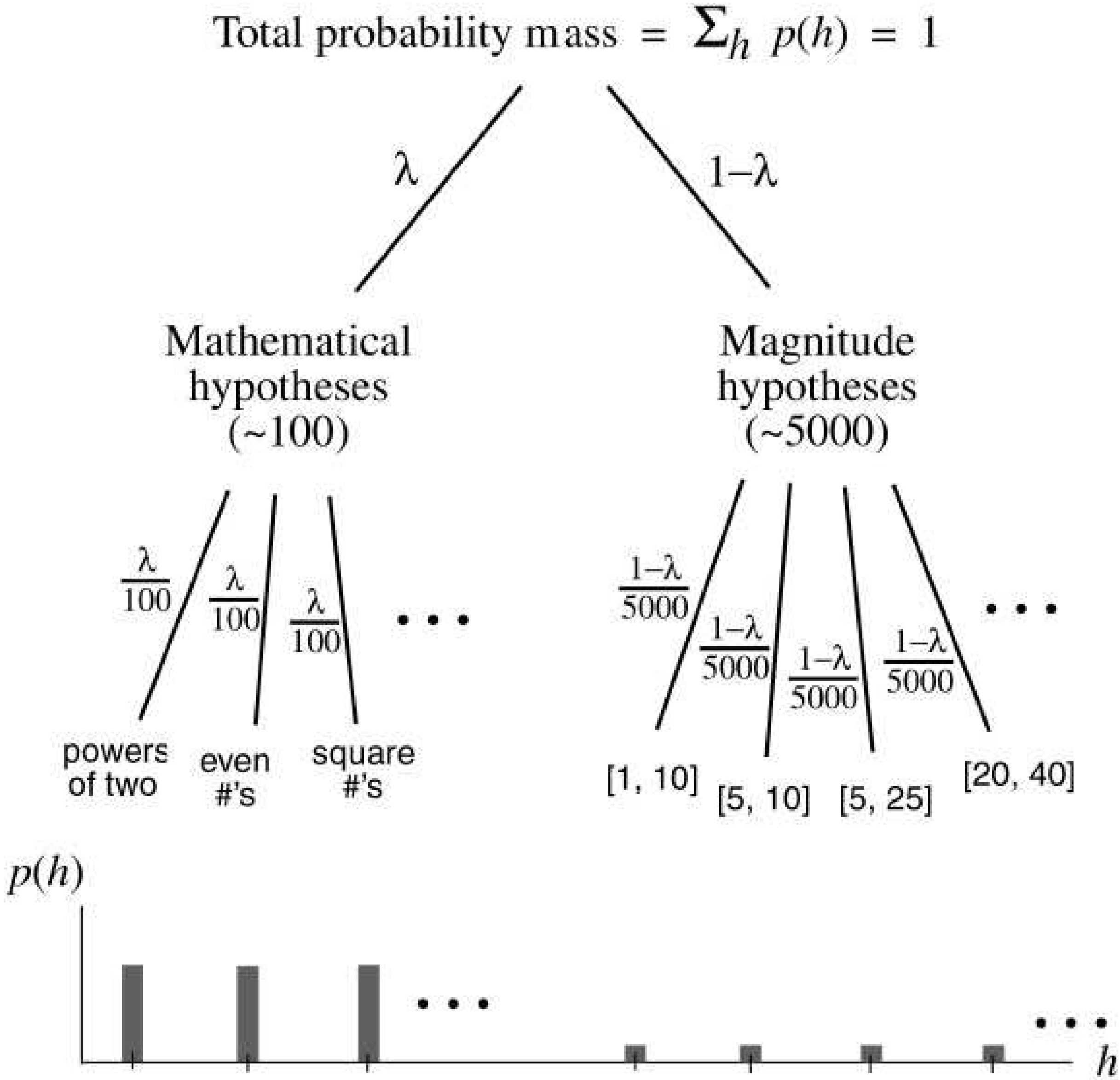
- Consider all $2^{2^2} = 16$ possible binary functions on 2 binary inputs

Boolean functions.

x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

- If we observe (x_1=0, x_2=1, y=0), this removes h_5, h_6, h_7, h_8, h_{13}, h_{14}, h_{15}, h_{16}
- Still leaves exponentially many hypotheses!
- Cannot learn efficiently without assumptions (no free lunch theorem)

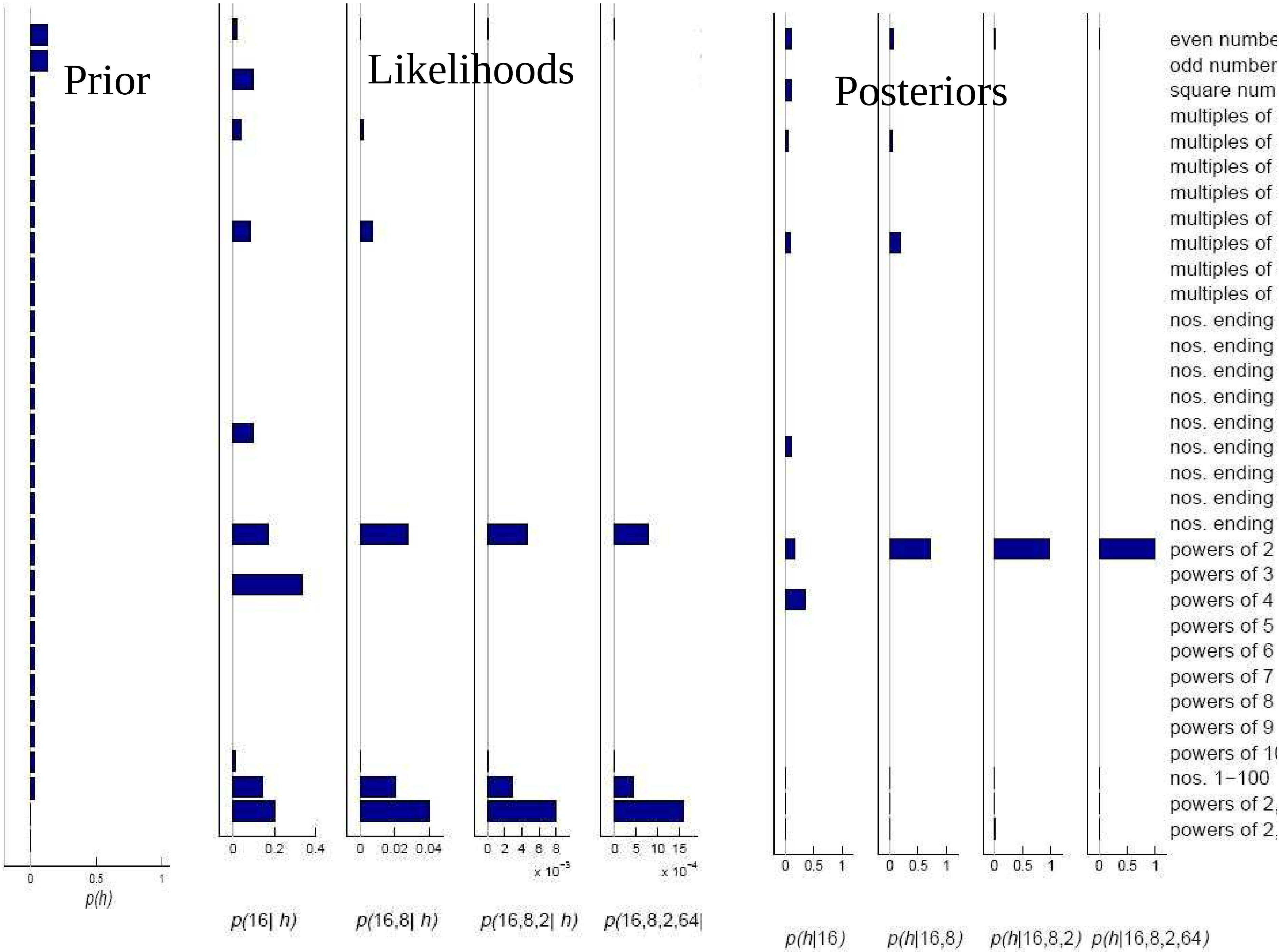
Hierarchical prior



Computing the posterior

- In this talk, we will not worry about computational issues (we will perform brute force enumeration or derive analytical expressions).

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$



Prior
Likelihoods
Posteriors
p(h)
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64|
p(h|16)
p(h|16,8)
p(h|16,8,2)
p(h|16,8,2,64)
even numbe
odd number
square num
multiples of
nos. ending
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 1(
nos. 1−100
powers of 2,

Generalizing to new objects

Given $p(h|X)$, how do we compute the probability that C applies to some new stimulus y?

$$p(y \in C \mid X)$$

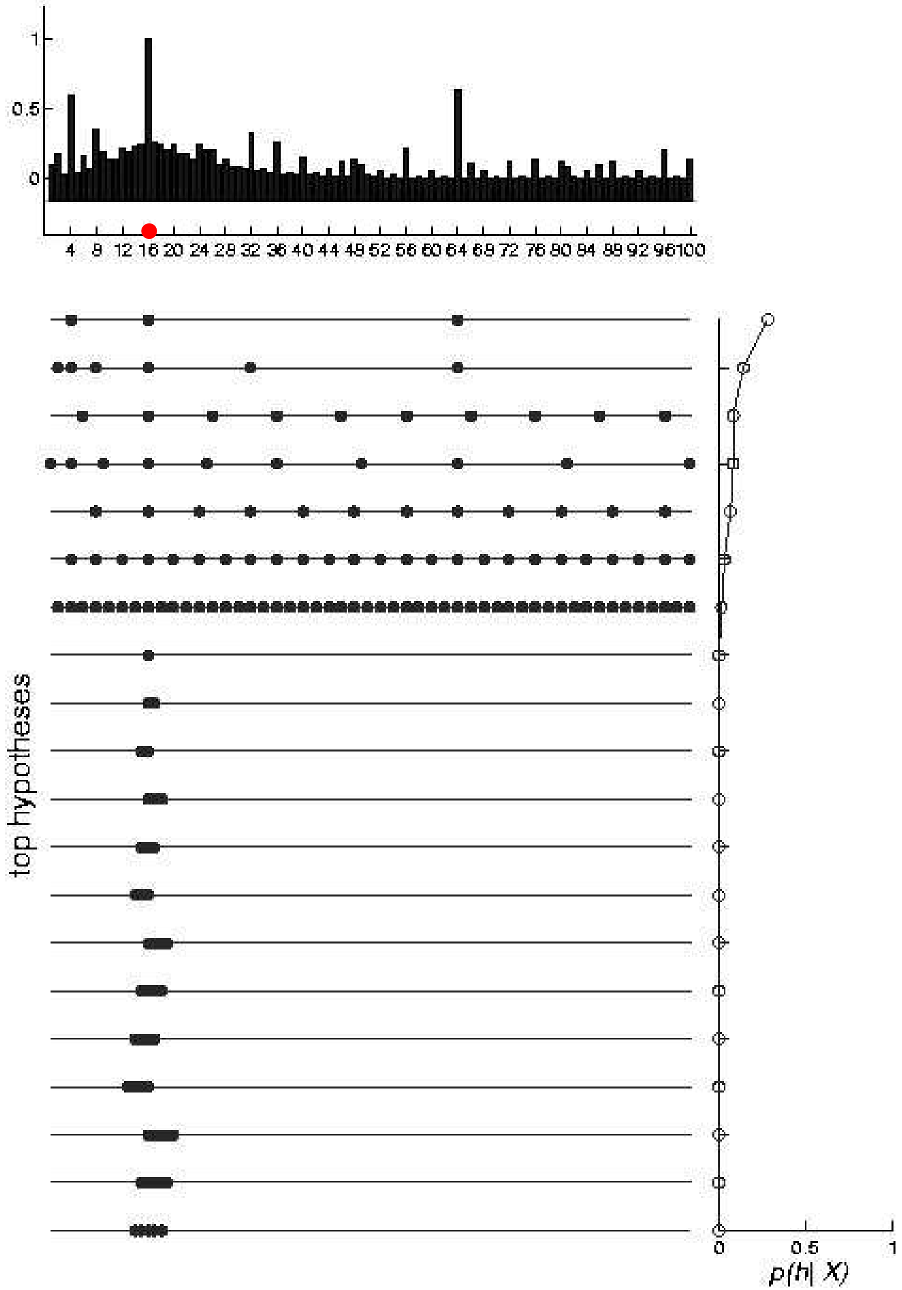
Posterior predictive distribution

Posterior predictive distribution

Compute the probability that *C* applies to some new object *y* by averaging the predictions of all hypotheses *h*, weighted by *p*(*h*|*X*) (**Bayesian model averaging**):

$$p(y \in C \mid X) = \sum_{h \in H} \underbrace{p(y \in C \mid h)}_{= \begin{cases} 1 \text{ if } y \in h \\ 0 \text{ if } y \notin h \end{cases}} \; p(h \mid X)$$

Examples:
16



1
0.5
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76
80
84
88
92
96
100
top hypotheses
0
0.5
1
p(h| X)

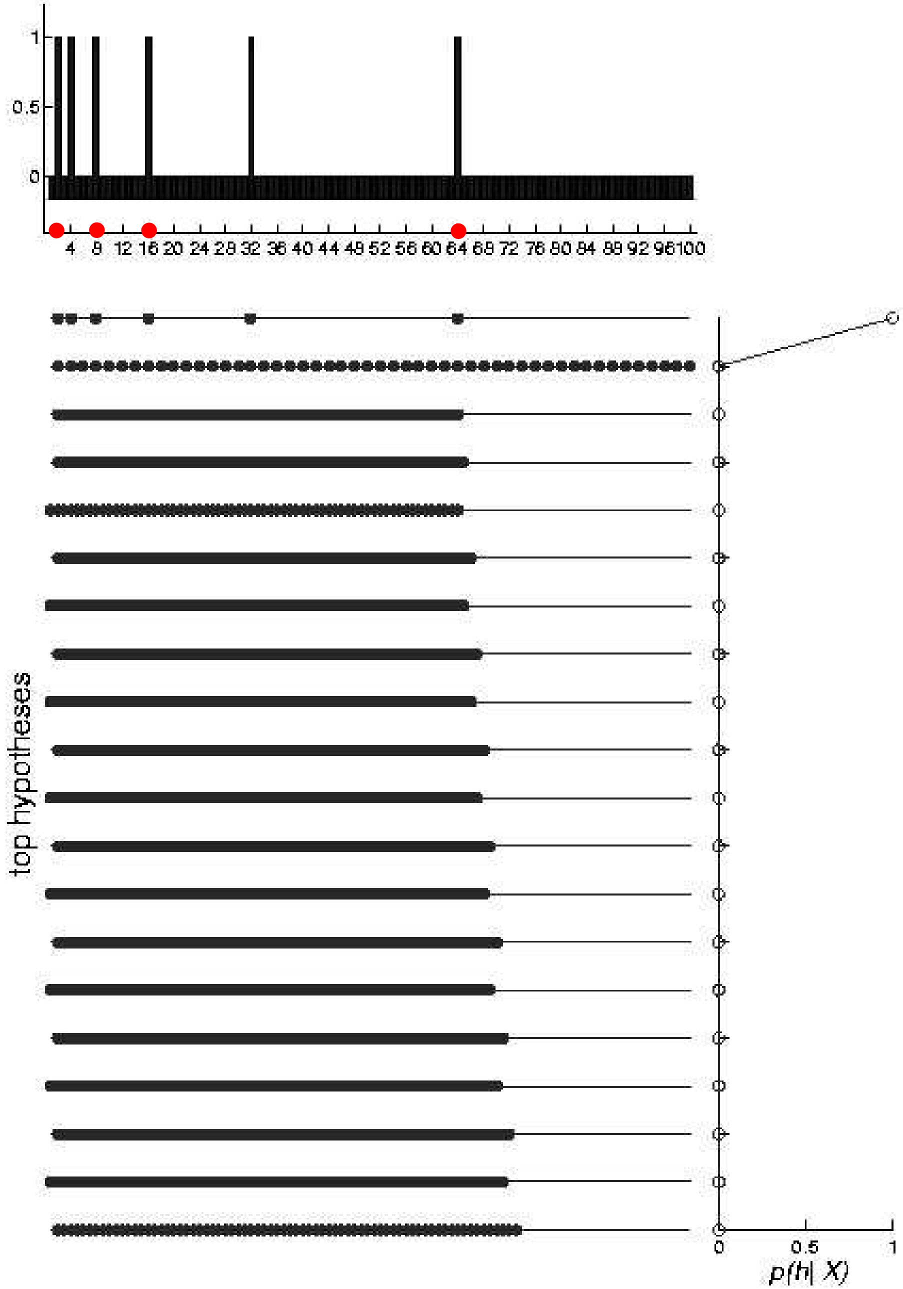
Examples:

16

8

2

64



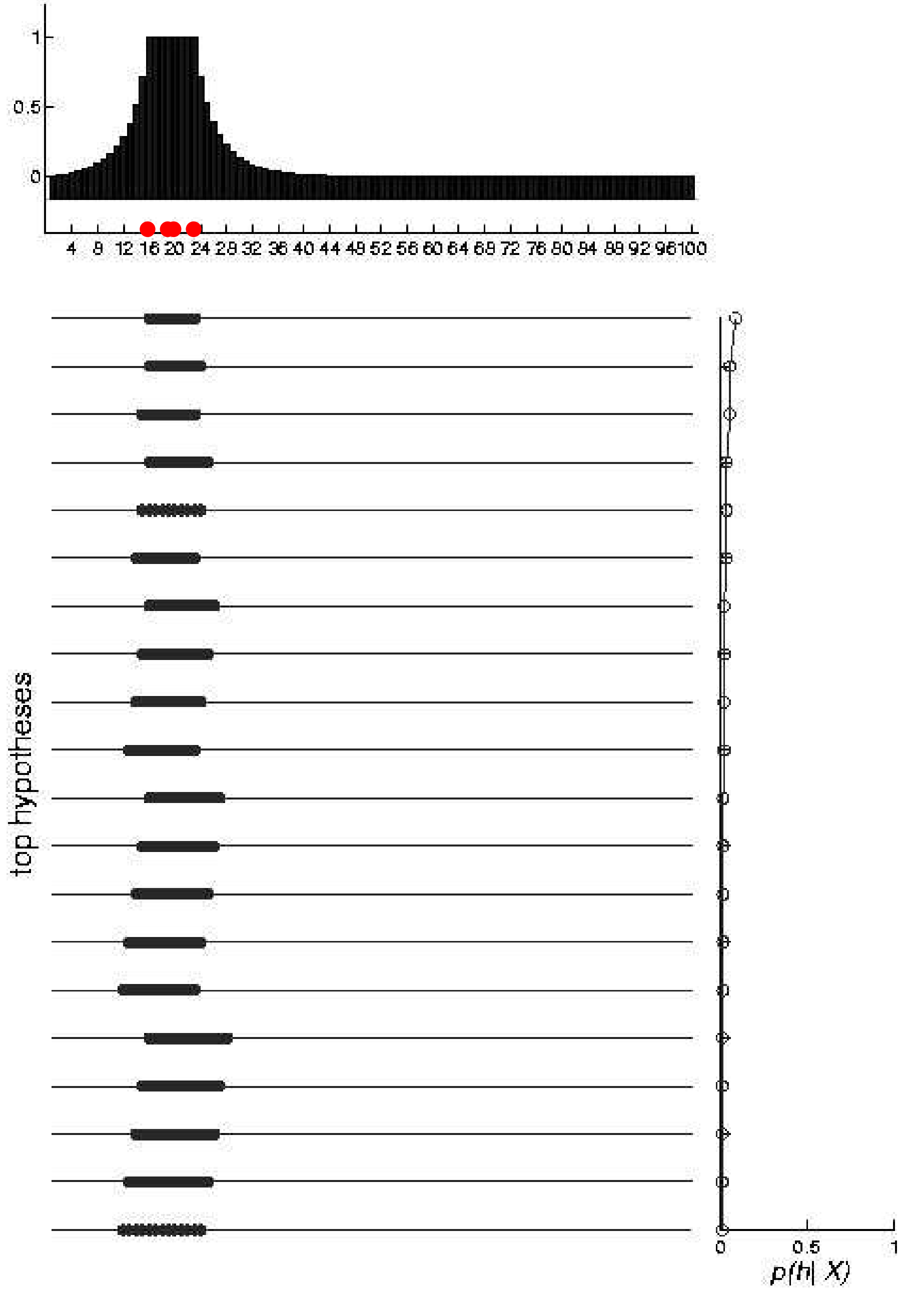
Examples:

16

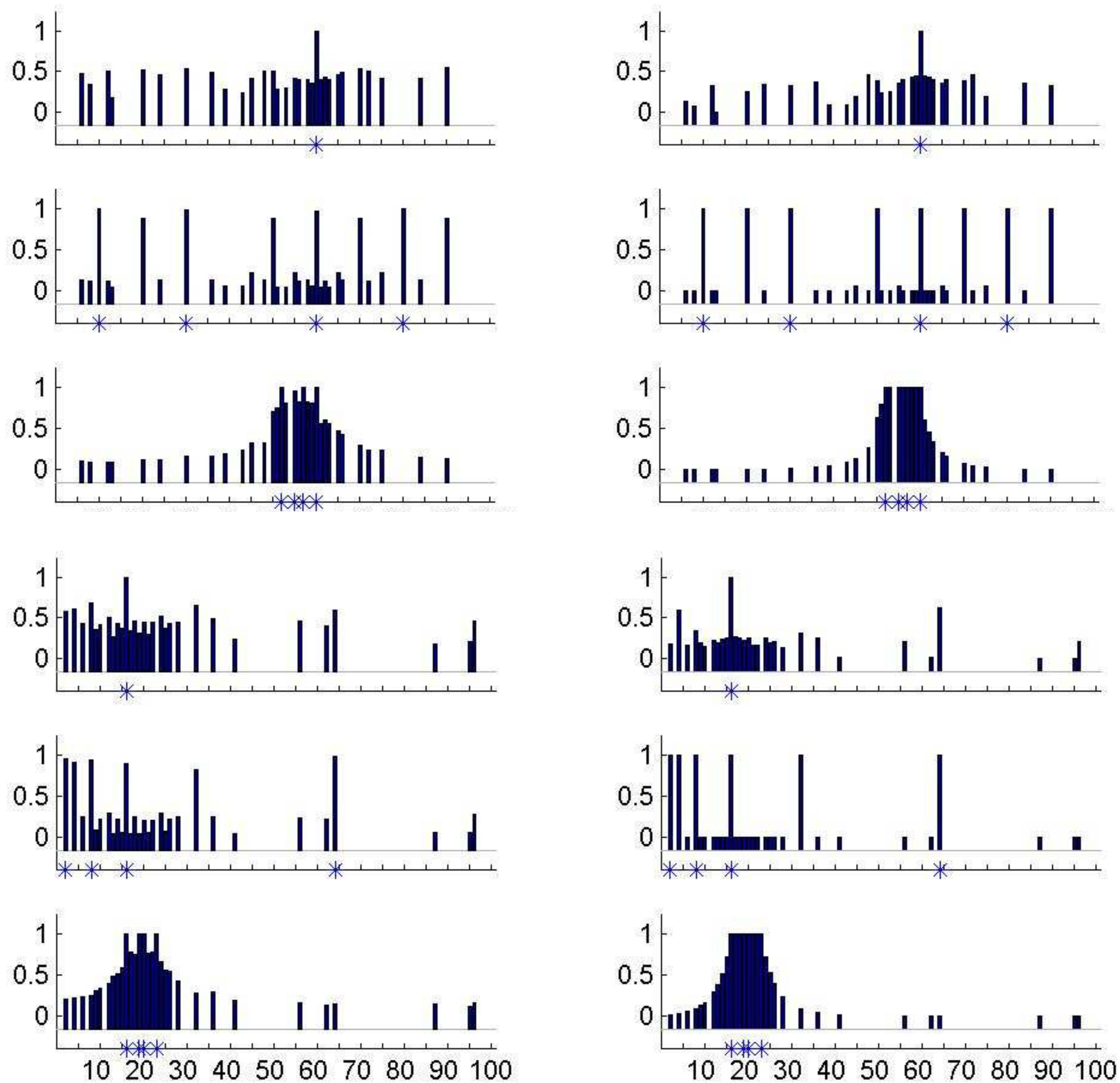
23

19

20



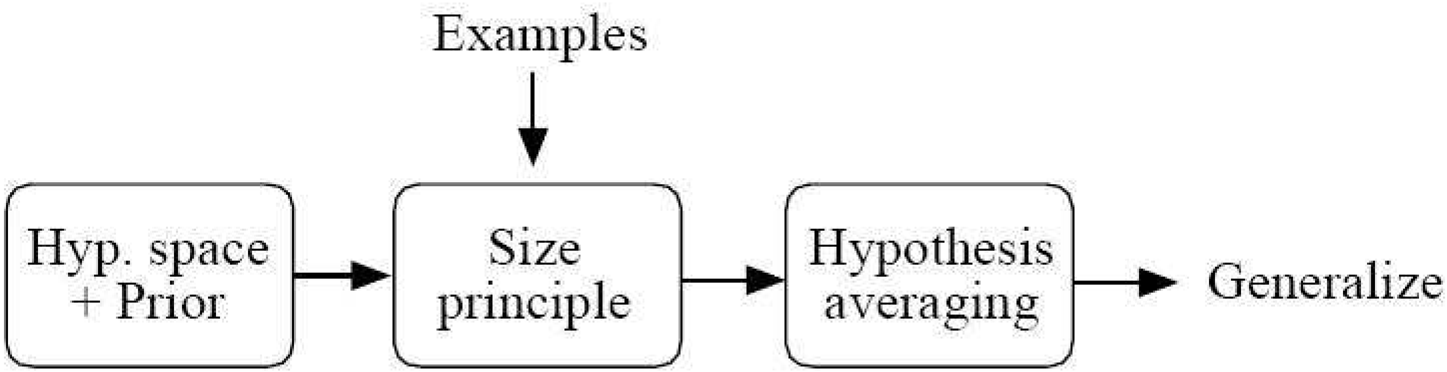
+ Examples	Human generalization	Bayesian Model
60		
60 80 10 30		
60 52 57 55		
16		
16 8 2 64		
16 23 19 20		



Rules and exemplars in the number game

- Hyp. space is a mixture of sparse (mathematical concepts) and dense (intervals) hypotheses.
- If data supports mathematical rule (eg X={16,8,2,64}), we rapidly learn a rule (“aha!” moment), otherwise (eg X={6,23,19,20}) we learn by similarity, and need many examples to get sharp boundary.

Summary of the Bayesian approach



1. Constrained hypothesis space H
2. Prior p(h)
3. Likelihood p(X|h)
4. Hypothesis (model) averaging:

$$p(y \in C | X) = \sum_h p(y \in C|h)p(h|X)$$

MAP (maximum a posterior) learning

- Instead of Bayes model averaging, we can find the mode of the posterior, and use it as a plug-in.

$$
\begin{aligned}
\hat{h} &= \arg\max_h p(h|X) = \arg\max_h p(X|h)p(h) \\
p(y \in C|X) &= p(y \in C|\hat{h})
\end{aligned}
$$

- As N $\rightarrow \infty$, the posterior peaks around the mode, so MAP and BMA converge

$$
p(y \in C|X) = \sum_h p(y \in C|h)p(h|X) \rightarrow \sum_h p(y \in C|h)\delta(h,\hat{h})) = p(y \in C|\hat{h})
$$

- Cannot explain transition from similarity-based (broad posterior) to rule-based (narrow posterior)

Maximum likelihood learning

- ML = no prior, no averaging.
- Plug-in the MLE for prediction:

$$\begin{aligned}\hat{h} &= \arg\max_h p(X|h) \\ p(y \in C|X) &= p(y \in C|\hat{h})\end{aligned}$$

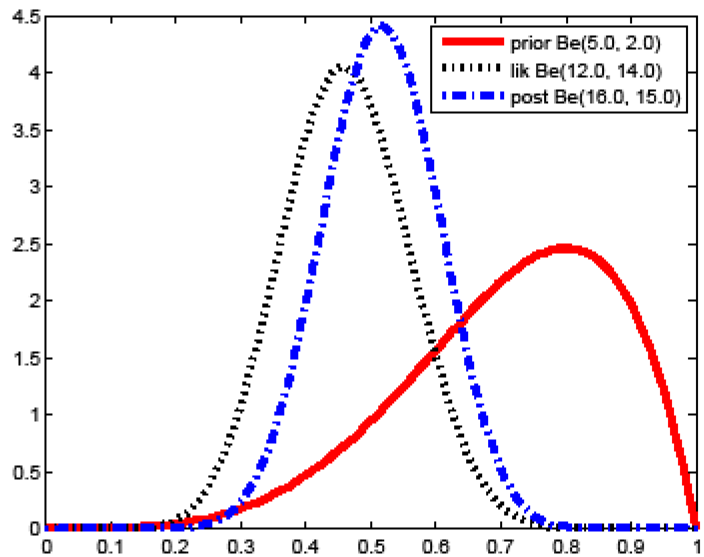
- X={16} -> h= "powers of 4" X={16,8,2,64} -> h= "powers of 2".
- So predictive distribution gets broader as we get more data, in contrast to Bayes.
- ML is initially very conservative.

Large sample size behavior

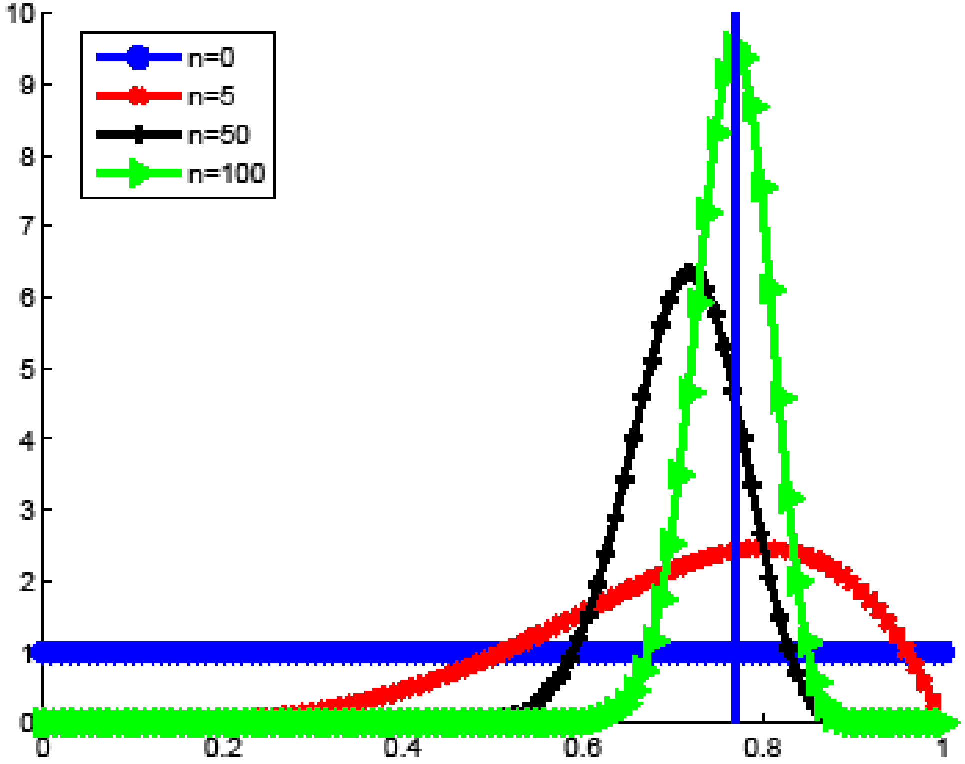
- As the amount of data goes to ∞, ML, MAP and BMA all converge to the same solution, since the likelihood overwhelms the prior, since $p(X|h)$ grows with N, but $p(h)$ is constant.
- If truth is in the hypothesis class, all methods will find it; thus they are consistent estimators.

Beta-Bernoulli model

$$
\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\
&= p(\mathcal{D}|\theta)\text{Beta}(\theta|\alpha_0, \alpha_1) \\
&= [\theta^{N_1}(1-\theta)^{N_0}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}] \\
&= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} \\
&\propto \text{Beta}(\theta|N_1+\alpha_1, N_0+\alpha_0)
\end{aligned}
$$



Sequential updating

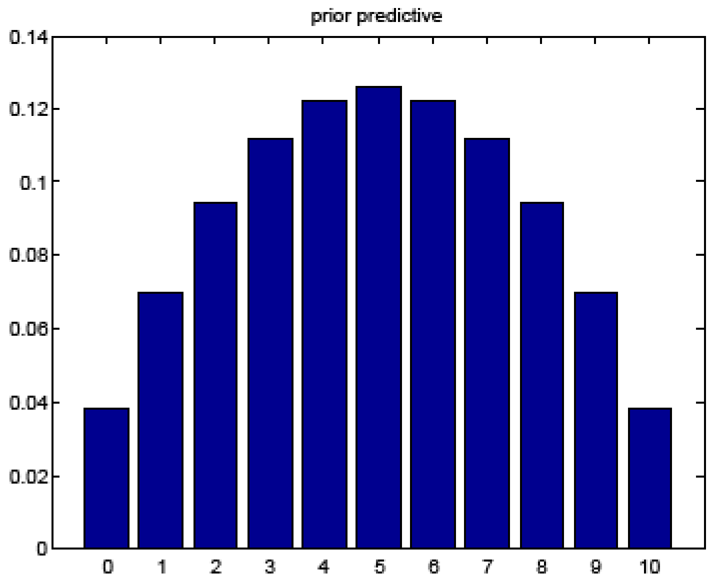


Prior predictive density

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

$$= \int_0^1 \text{Bin}(x|\theta, m)\text{Beta}(\theta|\alpha_0, \alpha_1)d\theta$$

$$\stackrel{\text{def}}{=} Bb(x|\alpha_0, \alpha_1, m) = \frac{B(x+\alpha_1, m-x+\alpha_0)}{B(\alpha_1, \alpha_0)} \binom{m}{x}$$



Posterior predictive density

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$

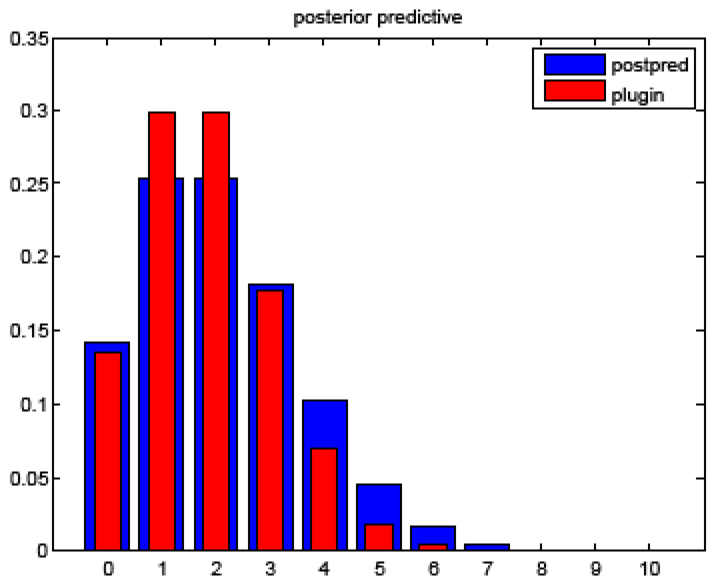
$$= \int_0^1 \text{Bin}(x|\theta, m)\text{Beta}(\theta|\alpha_0', \alpha_1')d\theta$$

$$\stackrel{\text{def}}{=} Bb(x|\alpha_0', \alpha_1', m) = \frac{B(x+\alpha_1', n-x+\alpha_0')}{B(\alpha_1', \alpha_0')}\binom{m}{x}$$

Plugin approximation

$$p(x|\mathcal{D}) = \int p(x|\theta)\delta_{\hat{\theta}}(\theta)d\theta = p(x|\hat{\theta})$$

$$= \text{Bin}(x|\hat{\theta}, m)$$



Posterior predictive

$$E[x] = m \frac{\alpha'_1}{\alpha'_0 + \alpha'_1}$$

$$\text{Var}[x] = \frac{m\alpha'_0\alpha'_1}{(\alpha'_0 + \alpha'_1)^2} \frac{(\alpha'_0 + \alpha'_1 + m)}{\alpha'_0 + \alpha'_1 + 1}$$

- If m=1, X in {0,1}, E[x|D] = p(x=1|D) = a1(a1+a0)

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})d\theta$$

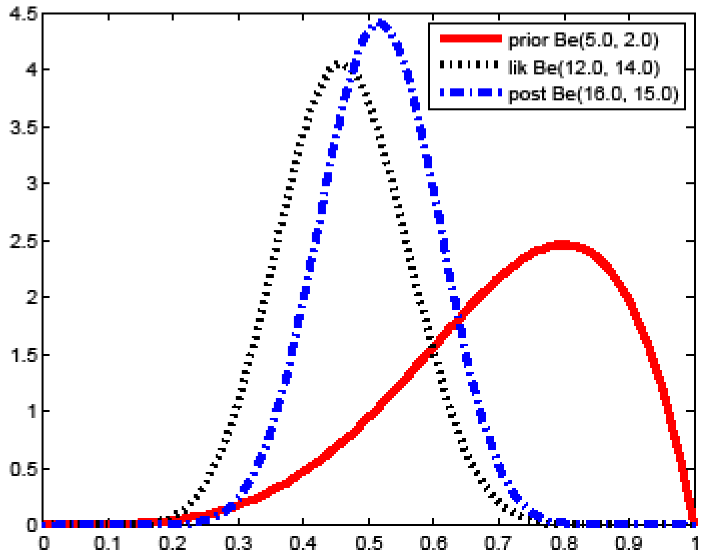
$$= \int_0^1 \theta \,\text{Beta}(\theta|\alpha'_1, \alpha'_0)d\theta = E[\theta|\mathcal{D}] = \frac{\alpha'_1}{\alpha'_0 + \alpha'_1}$$

Laplace's rule of succession

$$p(x = 1|\mathcal{D}) = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

Summary of beta-Bernoulli model

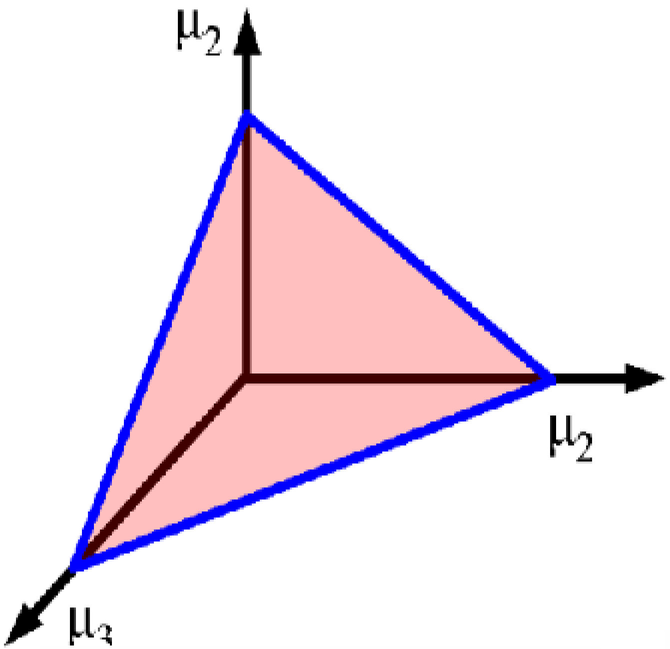
- Prior $$p(\theta) = \text{Beta}(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$
- Likelihood $$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$
- Posterior $$p(\theta|D) = \text{Beta}(\theta|\alpha_1 + N_1, \alpha_0 + N_0)$$
- Posterior predictive $$p(X=1|D) = \frac{\alpha_1 + N_1}{\alpha_1 + \alpha_0 + N}$$

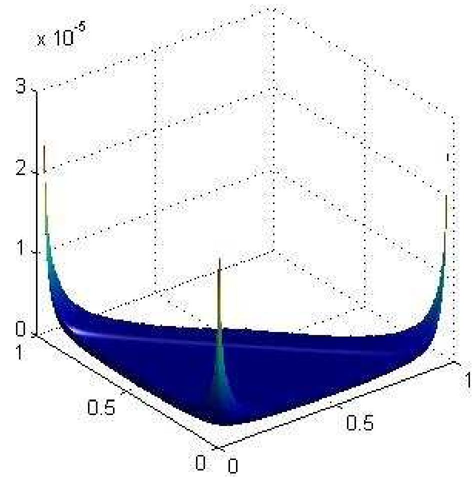


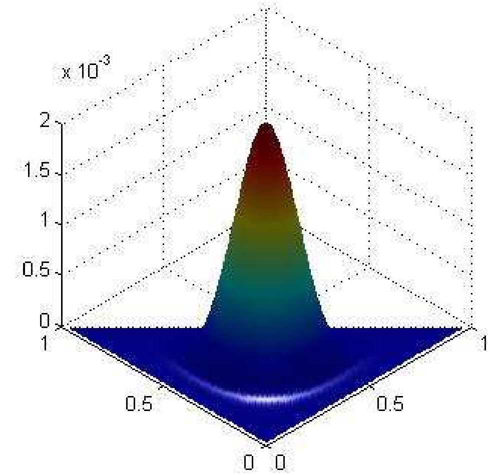
Dirichlet-multinomial model

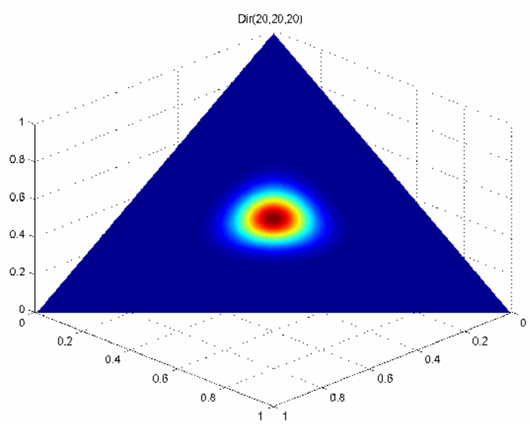
- $X_i \sim \text{Mult}(\theta,1)$, $p(X_i=k) = \theta_k$
- Prior $p(\theta) = \text{Dir}(\theta|\alpha_1, \ldots, \alpha_K) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
- Likelihood $p(D|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$
- Posterior $p(\theta|D) = \text{Dir}(\theta|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$
- Posterior predictive $p(X=k|D) = \dfrac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$

Dirichlet

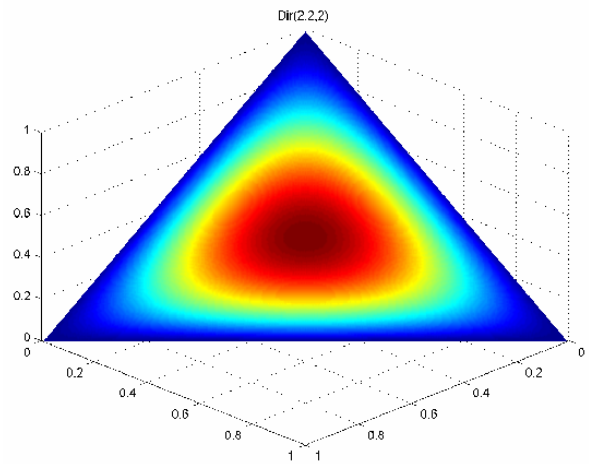




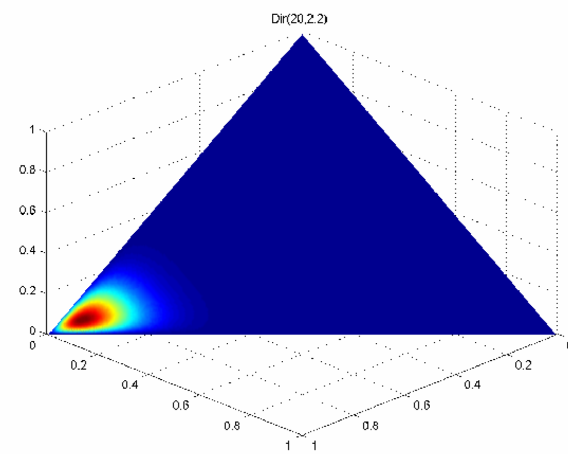




(20,20,20)



(2,2,2)



(20,2,2)

Summarizing the posterior

- If p(θ|D) is too complex to plot, we can compute various summary statistics, such as posterior mean, mode and median

$$
\begin{aligned}
\hat{\theta}_{mean} &= E[\theta|\mathcal{D}] \\
\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\
\hat{\theta}_{median} &= t : p(\theta > t|\mathcal{D}) = 0.5
\end{aligned}
$$

Bayesian credible intervals

- We can represent our uncertainty using a posterior credible interval

$$p(\ell \leq \theta \leq u | D) \geq 1 - \alpha$$

- We set

$$\ell = F^{-1}(\alpha/2), u = F^{-1}(1 - \alpha/2)$$

Example

- We see 47 heads out of 100 trials.
- Using a Beta(1,1) prior, what is the 95% credible interval for probability of heads?

```
S = 47; N = 100; a = S+1; b  = (N-S)+1; alpha = 0.05;
l = betainv(alpha/2, a, b);
u = betainv(1-alpha/2, a, b);
CI = [l,u]
 0.3749    0.5673
```